~~SECRET~~

THE CONTENT CONTROL CODE IN COMPUTER PROCESSING OF TELETYPES

June 1970

Central Intelligence Agency
Central Reference Service
Systems Analysis Staff

~~SECRET~~

Approved For Release 2002/10/09 : CIA-RDP80B01439R000500030004-2

1.0 Introduction

This paper responds to NSA's 9 June memorandum, Content Control Code Evaluation (Serial: N 0569), and recommends that NSA continue to use USIB Content Control Code (CCC) for at least a period of 12 months. The paper also supports NSA's request that other reporting units initiate experimental use of CCC. CRS believes that the content control code significantly enhances development of computer processing of teletype materials for dissemination, storage, indexing and searching, and that lack of the CCC, or a similar device, will seriously impair (if not curtail) near time progress in these areas.

The argument is derived from CRS's experiments with machine-aided dissemination during which some 20,000 COMINT messages have been processed by computer.

2.0 Background

Since the spring 1968, various experiments have been performed with machine-aided dissemination. The hypothesis was that since NSA teletypes were already in machine-readable language, and since dissemination processing represented a reasonably simple analytical function, that some practical results might be expected from machine-aided dissemination.

Although many computer text processing systems could have been chosen (e.g., one of those developed to support machine translation), the very available FMSAC-AIDDISSEM package was chosen for the early experiments. Since these experiments were designed for full Agency dissemination, rather than the more limited uses for which FMSAC-AIDDISSEM was developed, several system restrictions were soon surfaced. Accordingly, a new package was specified by CRS and built by the [redacted] 25X1A5a1
Presently this package is undergoing refinement within CRS. The first version of this new package, ALPHA-1, is now under test.

The summary of these developmental experiments is shown in the following table.

Table I Summary of Developmental Experiments

Date Span	Software	Volume of Teletypes	Number of Customer Offices	Dissemination Control
Spring 1968	AIDDISSEM	5200	7	Keyword
Winter 68/69	AIDDISSEM	8900	10	CCC
Spring 1969	Simulated SPEX-2	1300	39	"Two-Level"
	[redacted]	25X1A5a1		
Fall-Winter 69/70	SPEX-2	5000	up to 80	"Two-Level"

The "dissemination control" refers to those devices used to represent the user requirements in computer searching of teletype text:

Keyword:	Natural English language text words, or message externals.
CCC:	The Content Control Code—USIB 1967. Presently applied to some 45% of COMINT material.
CCC "Two-Level":	The Content Control Code used in conjunction with Keywords.

3.0 Summary of Conclusions

In general, machine-aided dissemination systems will operate at a high level of recall, but at the same time however, they can also be expected to cause some over-dissemination; that is, the impreciseness of the dissemination controls allows unwanted material to be disseminated; the results of the experiments discussed in Section 2 will be given in terms of this over-dissemination, and also in terms of three-levels of dissemination requirements complexity:

SECRET

1. Standard	A requirement calls for specific message series, summaries, or otherwise specifically-named message.
2. Area	A requirement calls for specific geographic area(s). A message is disseminated if it contains reference to the specific area.
3. Area/Subject	A requirement calls for a specific subject-concept within, between, or among named geographic areas.

Table II Over-Dissemination and Cumulative Volumes Disseminated

Requirement Complexity (Order-of-Increasing)	Keyword Over-Dissem.	CCC Two-Level Over-Dissem.	Cumulative Volume (Est.)
1. Standard	0%	Back-up Profiles	40%
2. Area	25%	5%	70%
3. Area & Subject	300%	25%	100%

Because a 300% over-dissemination is unacceptable, we conclude that, without the CCC two-level capability, dissemination of COMINT by computer text processing is presently limited to some 70% of the present dissemination volume—that is, to standard and area requirements; and further, that any retrospective searching of this material is not practically possible since this searching calls for area and subject complexity as the general rule.

The CCC is a powerful dissemination control because it represents originator/human analysis, in simple code notation, of message content. Thus, for example, the mere appearance of geographic area name "Keywords" in text is no guarantee that the message discusses substantively those geographic areas. The CCC area code would essentially guarantee it however.

Generally, the CCC represents a "lens" through which the message text can be viewed. On the one hand the contexts of English textwords are made more exact by a knowledge of the Area-Subject code which applies; also, the CCC area subject-code itself becomes more precise by coordinating it with words within the message. These are two examples of the "CCC Two-Level" approach.

There is another important advantage in using CCC as the dissemination control. Text processing systems become necessarily more sophisticated and expensive to operate as the number of words required to represent user requirements grow. In our experiment, thus far, we may show:

Table III Number of Terms Required to Represent User Requirements

	Standard	Area	Area & Subject (Est.)
Keywords	1000	4000	15,000 +
CCC Two-Level	100	400	1,000 (est.)

Thus, the "CCC Two-Level" approach requires significantly fewer words to represent the user requirements, decreases over-dissemination, and reduces computing costs measurably.

4.0 Present Status

Several projects are currently under way which are dependent upon further use of the CCC.

4.1 Project HERMES (Machine-Aided Dissemination Development)

Machine-aided dissemination of COMINT has begun, to a limited extent, within CRS. Presently we are utilizing the AIDDISSEM program on a pilot project, to disseminate some of the "standard" items to Agency-wide users.

We are also preparing dissemination profiles using the CCC and evaluating them by making test dissemination runs using the software developed for CRS [redacted] 25X1A5a1

We are also developing the ALPHA software which will have the capability of testing messages for the existence of the CCC, determining whether the CCC is valid, processing the message for dissemination if it is valid, and spilling the message out of the system if it is not.

4.2 Project INDIGO (Machine-Aided Indexing)

We are preparing for a six-month evaluation (from July-December 1970) of the use of the CCC two-level indexing approach. The ALPHA-1 software will be used to simulate the Machine-Aided Indexing; and the present human shallow indexing efforts will be used as an evaluation standard. [redacted] well known 25X1A5a1 for his documentation work, will be under contract for this six-month study.

4.3 Project EXTRA (Extract Tapes for RSM)

We are developing an operational capability for extracting from COMINT traffic messages which correspond to special information requirements; these messages are extracted to tape for further searching on the Rapid Search Machine.

4.4 Project CASTILE (Machine Storage of Teletypes)

We are developing a plan to replace the present storage of teletype printed copy with a machine-aided filing and/or storage system.

4.5 COMADS Monitor

This monitor is a computer program which analyzes incoming COMINT traffic according to countries and subjects discussed. The program writes statistical reports in terms of the code components contained within the reference serial notation and the Content Control Code notation appearing in the messages.